VLDB 2017

43rd International Conference on Very Large Data Bases

Reviews For Paper

Track	Research -> July 2016
Paper ID	278
Title	A General and Parallel Platform for Mining Co-Movement Patterns over Large-scale Trajectories

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Overall Rating	Accept
Summary of the paper (what is being proposed and in what context) and a brief justification of your overall recommendation. One paragraph	The paper presents a novel parallel algorithm for mining a general class of trajectory patterns. The paper is very easy to read containing all the details you need to understand it and the technique presented is efficient.
Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	1. Nicely written. This is one of the few papers I have ever reviewed that was clear from the beginning to the end. 2. Results are very useful and the algorithm presented makes the paper a potentally impactful one. 3. From the paper one can devise both the quality of the solution and also the reasons why it works. Every single decision is proved by a theorem.
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what are seen as negative aspect	I have not found any real mistake that it is worth being reported
Relevant for PVLDB	YES
Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium ratings for	Novel

delta papers and papers on well known topics but still with some valuable contribution).	
Significance	Improvement over existing work
Technical Depth and Quality of Content	Excellent work
Experiments	Very nicely support the claims made in the paper
Presentation	Excellent: careful, logical, elegant, easy to understand
Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	1. This is one of the few papers I've reviewed and I've understood with little (or no) effort 2. Results are very good and the presented experiments carefully evaluate all the possible aspects. 3. The newly proposed pattern generalizes those already presented in the literature 4. The parallel algorithm presented represents a very nice solution in practice given that the majority of solutions nowadays rely on very large clusters of inexpensive processing nodes running hadoop or spark. The main con that I've spotted is the fact that the algorithm might have been more nicely framed also within the Spark environment by taking advantage of the various possibilities offered by it (e.g. caching of RDDs)

Masked Reviewer ID: Assigned_Reviewer_2
Review:

Question	
Overall Rating	Weak Accept
Summary of the paper (what is being proposed and in what context) and a brief justification of your overall recommendation. One paragraph	The first contribution of the paper is the support of GCMP (General Co-Movement Pattern). This generalizes existing frameworks in putting a limit on the size of the gap, thereby mitigating the problem of loose-connection anomaly. The second contribution is the computation of all such patterns in a MapReduce framework. Going beyond a baseline scheme, the novelty is the idea of star partitioning and exploiting monotonicity in pruning.
Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	+ The analyses, in the form of various theorems, are nice. Even the baseline scheme with lemmas and theorem 1 shows that there is thoroughness in making sure that certain quality guarantees are provided. + The star partitioning and apriori enumeration are quite novel. Even though the algorithm itself is not novel, the development to make sure that monotonicity, in fact anti-monotonicity, is satisfied is solid work, i.e., section 5.2.1. + The experimental results are strong with three datasets.
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes,	- The GCMP generalization is not particularly novel. Putting a maximum gapsize on consecutive segments is well-known in sequence mining published more than 10 years ago. - In fact, I have doubts about formulating the GCMP patterns as proposed. Are

missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what are seen as negative aspect	we really interested in all sets of movements beyond a cardinality of size M? Take the Taxi dataset as an example. Let say that there are lots of taxis going from the airport to downtown. Let say that there are 1000 such taxis. For a given M, are we interested in 1000-chooses-M answers? So this speaks to the problem of picking M. If M is 500, what is 1000-chooses-500? In fact, even if the system gives the single answer of 1000-chooses-1000, I am not sure I am interested in this pattern as I already know that there are many taxis going from the airport to downtown. What I think I am really interested in are GCMP that are "anomalous", which is much harder to define.
Relevant for PVLDB	YES
Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium ratings for delta papers and papers on well known topics but still with some valuable contribution).	With some new ideas
Significance	Improvement over existing work
Technical Depth and Quality of Content	Solid work
Experiments	Very nicely support the claims made in the paper
Presentation	Excellent: careful, logical, elegant, easy to understand
Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	Regarding the second weak point, one line of related work is the super-imposition of constraints on spatio-temporal mining. An example is a road network. In other words, given a road network, the network imposes constraints on GCMP.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Overall Rating	Weak Accept
Summary of the paper (what is being proposed and in what context) and a brief justification of your overall recommendation. One paragraph	The paper proposes a general co-movement pattern (GCMP) that can model various patterns studied in previous works. Additionally, the proposed pattern also adds a temporal gap component to avoid loose-connection anomalies. The paper also investigate a scalable deployment of GCMP on big data platforms such as Hadoop (and Spark). In particular, both the implementations for a baseline sliding window, and a more sophisticated star-partitioning were provided. The experiments were conducted on three real-life data sets with detailed performance results. Overall, this is a solid paper, though it still could benefit from a few additional discussions on the Spark implementation as well as optimization details.
	S1. Provide a unified model that can represent major existing movement patterns, as well as addressing the issue with time gap in the group pattern. This

Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	is important and of high utility to the community as we can refer to a single framework for mining movement patterns. S2. Provide a MapReduce-based implementation (in Spark) for the framework, which is both scalable and efficient. S3. Combine various optimization strategies for apriori pruning with the star partition neighborhood. S4. The framework is open source.
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what are seen as negative aspect	W1. Though the implementation references Spark as the implementation platform (as it also reflects in github repo provided), the algorithm design is mostly limited to MapReduce, aka only Hadoop, which is a very small subset of Spark. This may have a negative impact on the baseline implementation. Particularly, recent releases of Spark have introduced window functions that can be applied directly in the sliding window scenario here. Certainly, the algorithm has to be redesigned to use DataFrame (and/or Spark SQL) interface, it has been noted that this is a very efficient way to execute window functions in Spark. W2. More details needed in the performance evaluation. Most implementations on Hadoop and Spark are very sensitive to data partitions, i.e. prone to data skewing issues. It seems that the (star-partition) implementation does not take this into account, and only use the default partitioning. It would be very helpful to have a discussion on this topic. In particular, it would be great to provide the difference in the number of partitions/splits, the amount of processing and memory usage (i.e. vcore and memory seconds) between TRPM and SPARE. W3. Both the star partitioning and apriori pruning are not new. What may be new is how they contribute to the big data implementation. This is not shown in the paper. A plot that breaks down the performance gain by each method would be greatly appreciated by the readers.
Relevant for PVLDB	YES
Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium ratings for delta papers and papers on well known topics but still with some valuable contribution).	With some new ideas
Significance	The paper is going to start a new line of research and products
Technical Depth and Quality of Content	Solid work
Experiments	OK, but certain claims are not covered by the experiments
Presentation	Reasonable: improvements needed

Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	Overall, this is a very nice-written paper, not only unifying (and improving) on the current model for movement patterns, but also providing an open-source framework that scales on big data platform. Though there are certain parts of the work that need more discussion and justification, this reviewer believes the research is of high interest to the community and could start a new direction in supporting trajectory mining at scale. Thus, I would lean toward an accept. Other details: - Some choices of words may need to be reconsidered: for example, "a bunch of" might not be appropriate in a technical paper. - References to star partitioning and apriori pruning are missing. Though these are well-known, they need to clearly cited. At least the following reference is missing: + Yoo, J. and Shekhar, S. "A Join-less Approach for Mining Spatial Co-location Patterns". IEEE TKDE 2006. - In "In contrast, when utilizing the multi-core environment, SPARE-P achieves 7 times speedup and SPARE-S achieves 10 times speedup.", was "multi-core" referring to the use of all 16-cores in one of your node? The specification of the machine was not clear. - In the source-code available on github, in: + src/main/java/kreplicate/KReplicateLayout.java The computation of "eta" was slightly different than that in the paper: " eta = ((int) (Math.ceil(k*1.0/l)) - 1) *g + K+L-1;" Note this is times "g", and not "(g-1)" as shown in the formula. It would be great for this to be consistent in both.